DOCUMENT RESUME

ED 430 987 TM 029 813

AUTHOR	Falk, Beverly; Ort, Suzanna Wichterle; Moirs, Katie
TITLE	Using Classroom-Based Assessment on a Large Scale: Supporting and Reporting on Student Learning with the Early Literacy Profile.
INSTITUTION	Columbia Univ., New York, NY. Teachers Coll. National Center for Restructuring Education, Schools and Teaching.
PUB DATE	1999-00-00
NOTE	49p.; Paper presented at the Annual Meetings of the Association for Supervision and Curriculum Development (San Francisco, CA, March 6-9, 1999) and the American Educational Research Association (Montreal, Quebec, Canada, April 19-23, 1999).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Academic Standards; Educational Assessment; Educational Trends; *Emergent Literacy; *Performance Based Assessment; Primary Education; Profiles; Research and Development; State Programs; *State Standards; Test Reliability; Test Validity; Testing Programs
IDENTIFIERS	Large Scale Programs; *New York

ABSTRACT

This paper describes the development work and research findings of an initiative to create a statewide literacy assessment in New York to inform teaching and learning and report on group performance trends. The Early Literacy Profile (ELP) is a classroom-based, standards-referenced performance assessment for students in the primary grades organized around four purposes for language use: information and understanding, literary response and expression, critical analysis and evaluation, and social interaction. Studies were conducted to see how well the ELP meets these purposes. In 1997-98, 63 teachers representing 19 schools piloted the ELP with approximately 1,215 students in grades 1 to 3. The ELP was evaluated for construct validity, content validity, student performance, and criterion validity. These evaluations found the ELP to be a valid assessment of literacy progress that is technically strong in that it effectively discriminates levels of student performance and can be scored reliably. Evaluations also found the ELP to be instructionally useful. One of the most powerful findings of the studies was the degree to which teachers reported the ELP to be supportive of their teaching and their students' learning. (Contains 9 tables and 58 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 430 987

Using Classroom-Based Assessment on a Large Scale: Supporting and Reporting on Student Learning with The Early Literacy Profile

Paper Presented at the Annual Meeting of the American Educational Research Association, April 1999, in Montreal, Canada

Beverly Falk
and
Suzanna Wichterle Ort
National Center for Restructuring Education, Schools, and Teaching (NCREST), Teachers College, Columbia University

with
Katie Moirs, Connecticut State Education Department

TM029813

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
S. Wichterle Ort
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This paper describes the development work and research findings of an initiative to create a state-wide literacy assessment that can connect to and inform teaching and learning as well as report on group performance trends. Designed and researched for the New York State Education Department by the National Center for Restructuring Education, Schools, and Teaching (NCREST) at Teachers College, Columbia University in collaboration with New York State teachers, the Early Literacy Profile (ELP) is a classroom-based, standards-referenced, performance assessment for students in the primary grades. What follows is a description of the Profile, an explanation of its theoretical underpinnings, an account of pilot studies conducted in 1997-1998, a discussion of study findings, and recommendations/questions for further research and development.

PART I: OVERVIEW

The ELP is an assessment designed to provide information about student progress in various aspects of literacy development - reading, writing, speaking, and listening. It is organized around four purposes for language use as outlined in the New York State Learning Standards for the English Language Arts: 1) information and understanding, 2) literary response and expression, 3) critical analysis and evaluation, and 4) social interaction.

The ELP consists of a small set of standardized tasks that are to be completed in the context of classroom life, collected at designated times in the year, and evaluated in relation to developmental scales. Student proficiencies are assessed by examining the following sets of evidence:

Please do not cite, reproduce, or distribute without permission of authors

Reading Evidence

- **Reading Sample**: teacher's documented observation of a student's reading that analyzes oral reading fluency and comprehension (See Figure 1)
- **Reading List**: list of texts that each student has that provides evidence about the student's range and experience as a reader
- **Reading Response**: student's written response to a text that provides additional information about the student's abilities to understand and analyze texts

Writing Evidence

- **Story/Narrative - First Draft**
- **Same Story/Narrative - Final Draft**
- **Reading Response**: same as used in the Reading Evidence section but used here to provide information about the student's independent writing abilities

Listening/Speaking Evidence

- Teacher's documented observations of a student engaged in speaking and listening for social interaction

Diagnostic Tools

- **Alphabetic Principle task**
- **Phonemic Awareness task**
- **Word Recognition task**

Based on their evaluation of the various pieces of evidence, teachers assign students scale scores along a continuum of progress in reading, writing, and listening/speaking. The dimensions described in the scales are key components of preparation for achievement of the New York State Elementary English Language Arts standards. Reading and Writing scales have 4 major stages, subdivided into 8 scale points (see Figure 2). Each scale point corresponds to a number:

Falk & Ort, 1999
Please do not cite, reproduce, or distribute without permission of authors

FIGURE 1.

Reading Sample

Name: _____ Grade: _____ **Side 1**

Fluency

Date	FALL (__/__/__)		SPRING (__/__/__)	
Title/Author				
Type of Text (Each type of text correlates to a scale point on the reading scale)	1. Simple picture book 2. Patterned language / picture book 3. Easy beginning reader / short chapter / content area book 4. Beginning reader / short content area book 5. Easy chapter / content area book 6. Medium level chapter /content area book 7. Challenging children's literature 8. Complex children's literature		1. Simple picture book 2. Patterned language / picture book 3. Easy beginning reader / short chapter / content area book 4. Beginning reader / short content area book 5. Easy chapter / content area book 6. Medium level chapter /content area book 7. Challenging children's literature 8. Complex children's literature	
	Rate	**Comments:**	**Rate**	**Comments:**
Engaged in pretend reading	1 2 3 4		1 2 3 4	
Used book language	1 2 3 4		1 2 3 4	
Told the story from the pictures	1 2 3 4		1 2 3 4	
Understood directionality of print	1 2 3 4		1 2 3 4	
Focused on print	1 2 3 4		1 2 3 4	
Displayed 1:1 correspondence	1 2 3 4		1 2 3 4	
Drew on previous experience to make sense of text	1 2 3 4		1 2 3 4	
Used pictures to aid understanding	1 2 3 4		1 2 3 4	
Recognized high frequency words in text	1 2 3 4		1 2 3 4	
Had a substantial sight vocabulary with few miscues	1 2 3 4		1 2 3 4	
Used graphophonic strategies: Beginning consonants / root words / endings / syllabication	1 2 3 4		1 2 3 4	
Used syntactic strategies: Predictable language patterns / sentence and grammatical structure	1 2 3 4		1 2 3 4	
Used semantic strategies: picture clues / recall of story line	1 2 3 4		1 2 3 4	
Used a variety of cueing systems	1 2 3 4		1 2 3 4	
Was independent in problem solving text difficulties	1 2 3 4		1 2 3 4	
Monitored own reading	1 2 3 4		1 2 3 4	
Self corrected	1 2 3 4		1 2 3 4	
Projected meaning with oral expression	1 2 3 4		1 2 3 4	
Maintained momentum and fluency	1 2 3 4		1 2 3 4	
Demonstrated confidence	1 2 3 4		1 2 3 4	

Rating Key 1=Not at All 2= A Little 3= Moderately 4= A Lot If a choice does not apply, write N/A (not applicable) in the "comments" space

BEST COPY AVAILABLE

Reading Scale

FIGURE 2.

Name: _____

	Emergent Reader		Beginning Reader	
	1 Early Emergent	**2** Advanced Emergent	**3** Early Beginning	**4** Advanced Beginning
Type of text that student reads at instructional level	**Simple picture books:** • Text generally focuses on a single idea. • Words in text are large, well-spaced with only one or two lines on each page. • Words or sentence patterns in text are repeated. • A simple illustration or picture accompanies and directly corresponds to the text. • Texts are brief ranging from 10-36 words. • Examples of these kinds of books are The Birthday Cake and Colors.	**Simple texts/Patterned language/picture books:** • Text may contain rhyming, repetitive words, phrases, and actions in predictable language structures. • Pictures provide much of the main idea. • Texts range from 25-70 words. • Examples of these kinds of books are My Garden and Fun with Hats.	**Complex patterned language books/Easy beginning reader/ short chapter/content area books:** • Illustrations provide important information about the text. • There are many high frequency words. • Text begins to introduce simple plot, theme, problems, and/or solutions. • Repetition of events and/or words is in each of the stories. • Natural language patterns begin to appear. • Text generally consists of several lines per page. • Examples of these kinds of books are Brown Bear and Go Dog Go.	**Beginning reader/ short chapter/content area books:** • Illustrations support the text. • There are many high frequency words. • Text has simple theme, plot, and problems and/or solutions. • Some repetition of events in each of the stories. • Natural language patterns are incorporated. • Text generally consists of several lines per page. • Examples of these kinds of books are Nosy Nora and More Spaghetti I Say.
Strategies	• Begins to display awareness of some conventions of print: front-back of book, left-right progression, top-bottom. • Focuses mainly on pictures. • Begins to know difference between pictures and print. • May recognize familiar words occasionally. • May engage in "pretend" reading. • Needs much assistance and support from an adult.	• Awareness of conventions of print is established: front-back of book, left-right progression, top-bottom. • Focuses on print as well as pictures. • Knows difference between pictures and print and begins to demonstrate awareness of 1:1 correspondence. • Knows a few familiar words on sight. • May engage in "pretend" reading, using book language ("once upon a time," "the end," etc.). • May "read" from memory. • May draw upon predictable language patterns to anticipate or recall the text. • May demonstrate awareness of some letter-sound relationships - mostly beginning consonants. • Needs assistance and support from an adult.	• Focuses mostly on the print to decipher meaning. • Picture clues still provide crucial support. • 1:1 correspondence is fairly established. • Small sight-word vocabulary is established. • Some high frequency words in text are recognized. • Begins to use some aspects of cueing systems: letter-sound relationships (especially beginning and ending consonants, root words), recall of story line, and predictable language structures. • Starts to monitor and self-correct own reading. • Reading requires considerable effort and continued assistance from an adult.	• Focuses on print to decipher meaning, refers to pictures only for support. • Stable sight vocabulary is established. • High frequency words are recognized. • Uses wider range of cueing systems while reading, sometimes with prompting from adult. • Monitoring and self-correcting are becoming established. • Reading requires some effort and assistance from an adult.
Comprehension	• Uses pictures almost exclusively to gain understanding. • Can talk about the pictures.	• Relies mostly on pictures clues and recall of story line to make sense of print. • Demonstrates understanding of a story through comments, reactions, discussion, and/or drawings.	• Relies more on print to find the meaning of text, but pictures play an important role. • Can identify some of the attributes in the text: character, setting, problem, resolution. • Retells general story or subject of text, but awareness of sequence of events may not be demonstrated.	• Relies mainly on print to find meaning of text with pictures playing a supporting role. • Can identify and discuss many of the attributes in the text: main idea or subject, characters, setting, problem and resolution. • Retells general story or subject of text with some awareness of sequence of events.

	Independent Reader		Experienced Reader	
	5 Early Independent	**6** Advanced Independent	**7** Experienced	**8** Very Experienced
Type of text that student reads at instructional level	**Easy chapter/content area books:** • Texts may have illustrations that provide moderate to minimum support. • Text has clearly evident theme or plot with conclusions. • Literary language structures are integrated with natural language. • Examples of these kinds of books are Nate the Great to Madeline.	**Medium level chapter/content area books:** • Texts have illustrations that provide minimal to incidental support. • Increasingly complex information, plot line, character or relationships. • Literary language structures are integrated with natural language. • Examples of these kinds of books range from The Stories Julian Tells to Freckle Juice.	**Challenging children's literature:** • Occasional pictures enhance the story. • Language structures and vocabulary are increasingly complex. • Background knowledge and higher-level thinking skills may be needed to understand and appreciate humor, problem, or suspense. • Chapters build upon each other. • Text size is usually a bit smaller. • Examples of these kinds of books range from Amber Brown is not a Crayon to Little House on the Prairie.	**Complex children's literature:** • Texts may not have any illustrations. • Texts address issues and/or information with shades of meaning requiring interpretation and analysis. • Imagery and metaphors may be used. • Chapters may move in time, place, and perspective. • Examples of these kinds of books range from Charlie and the Chocolate Factory to Roll of Thunder, Hear My Cry.
Strategies	• Begins to use cueing systems to problem-solve text difficulties with increasing independence. • Substantial sight vocabulary is established. • Monitoring and self-correcting are used regularly. • Oral reading begins to have some momentum. • Expression begins to be evident in oral reading. • Developing confidence and independence as a reader.	• Uses all cueing systems independently to problem-solve occasionally encountered text difficulties. • Recognizes most words in text. • Monitoring and self-correcting are used regularly and quickly to solve problems. • Oral reading has increasing momentum and expression with attention to punctuation. • Demonstrates increasing confidence and independence as a reader.	• Cueing systems are mostly internalized. • Few word problems encountered during reading. • Problem-solving strategies are mostly internalized. • Oral reading has momentum and fluency. • Effective oral expression and attention to punctuation. • Demonstrates confidence and independence.	• Cueing systems are almost entirely internalized. • Problems are encountered rarely during reading. • Problem-solving strategies are internalized. • Oral expression during reading reflects the subtleties of mood, pace, and tension of text. • Demonstrates strong sense of confidence and independence. • Reading pace and stamina is strong.
Comprehension	• Can identify and discuss main idea or subject, characters, setting, problem and resolution. • Retells and discusses story or subject of text with awareness of sequence of events.	• Can identify and discuss main idea or subject, characters, setting, problem resolution. • Retells and discusses subject of text with awareness of sequence of events and attention to relevant details.	• Can provide a coherent synthesis of the meaning of the text, using relevant details. • Analysis and interpretation skills are developing. • Awareness of the literary aspects of text and/or nuances and subtleties of text is developing. • May relate ideas in text to other ideas, experiences, and/or literature when prompted.	• Can provide a coherent synthesis of the meaning of the text, with extensive use of relevant details. • Analysis and interpretation skills are clearly evident. • Appreciates nuances and subtleties of text. • Has strong awareness of literary aspects of text. • Can volunteer independently how ideas in text relate to other ideas, experiences, and/or literature.

BEST COPY AVAILABLE

Major Stage	Scale Points
Emergent Reader/Writer	1=Early Emergent Reader/Writer 2=Advanced Emergent Reader/Writer
Beginning Reader/Writer	3=Early Beginning Reader/Writer 4=Advanced Beginning Reader/Writer
Independent Reader/Writer	5=Early Independent Reader/Writer 6=Advanced Independent Reader/ Writer
Experienced Reader/Writer	7=Experienced Reader/Writer 8=Very Experienced Reader/Writer

The Listening/Speaking scale describes 4 stages -- Emergent, Beginning, Independent, Experienced.

An additional section of the Profile contains three diagnostic tools that can be used to take a deeper look at the progress of students who are in early stages of literacy learning. These assessment tasks examine students' grasp of some important skills - the alphabetic principle, phonemic awareness, and word recognition - that recent research reports suggest are essential for effective and fluent reading and writing (International Reading Association and the National Association for the Education of Young Children, 1998; National Research Council, 1998).

All work reported on in this paper refers to the first three sections of the ELP. No studies have been conducted on the Diagnostic Tool Section, which yields student scores separate from the Reading, Writing, and Listening scales.

BEST COPY AVAILABLE

Please do not cite, reproduce, or distribute without permission of authors

Influences on Profile Development

The ELP was created with the input of many New York State educators. The development team, led by NCREST, included hundreds of elementary school teachers who participated in several pilots of the Profile since 1996. Also involved were New York State Education Department associates and consultants from the Center For Educational Options at the City University of New York and the Educational Testing Service in Princeton, New Jersey.

The ELP has also been informed by and adapted from other existing early literacy assessments:

- The Primary Language Record/California Learning Record (Barrs et al., 1988)
- The Reading/Writing Scale of the South Brunswick, New Jersey Public Schools (South Brunswick, New Jersey Public Schools, 1992)
- The Student Outcomes and Developmental Stages of the Rochester, New York Public Schools (Rochester, New York Public Schools, 1993)
- The American Literacy Profile Scales (Griffin, Smith & Burrill, 1995)
- "First Steps" Developmental Continuums of the Education Department of Western Australia (Education Department of Western Australia, 1994)

Purposes of Profile Use

The ELP aims to meet the challenge of finding a literacy assessment for the early elementary grades that simultaneously serves several purposes: supporting learning, informing instruction, and being useful for accountability purposes. It is designed to:

- Prepare students in the primary grades to meet the elementary level of the New York State English Language Arts standards;
- Demonstrate student progress over time to teachers, students, and their families;

Falk & Ort, 1999
Please do not cite, reproduce, or distribute without permission of authors

- Build teachers' capacities to support students' literacy development and their progress toward the standards;

- Identify students who require extra supports or intervention;

- Provide information about group performance to help administrators and policymakers make decisions about where resource allocations are most needed.

The ELP is conceived as an instructional assessment, providing a link between standards and instruction for accountability purposes.

Theoretical Underpinnings

The ELP can best be understood by explaining key issues that have been addressed in the design of the instrument. This section of our paper addresses three main points: 1) characteristics of young children's learning and how the ELP embodies this knowledge; 2) essential elements of literacy and how these have guided ELP development; and 3) technical demands of large-scale assessment and how the ELP has taken them into account.

1. Teaching and Testing the Way Young Children Learn

Cognitive research over the last several decades has led to deeper understandings about the nature of the learning process. Three important ideas from this research are incorporated into the design of the ELP. The first is that learners acquire information and develop concepts through active interaction with a range of experiences, ideas, and relationships. This process of learning is not linear -- with "basic" skills preceding thinking skills -- but rather, is supported best when skills are combined with higher-order thinking, embedded in contexts, and applied to real-world situations (Bruner, 1960; Falk, 1996; Fosnot, 1989; National Association for the Education of Young Children, 1988; Piaget & Inhelder, 1970; Resnick, 1987; Sternberg, 1985; Vygotsky, 1978). Informed by this view of learning, the ELP examines the literacy learning process in as close an approximation of the natural learning environment as technical constraints allow. It documents students' literacy

Please do not cite, reproduce, or distribute without permission of authors

development directly through actual performance in the context of classroom life rather than in tests that indirectly evaluate literacy. Students demonstrate their reading skills by reading with the teacher and discussing the text; they demonstrate their writing skills by completing written pieces composed in response to purposeful assignments; and they demonstrate their listening/speaking skills by engaging in discussions with their classmates.

The second understanding about learning that has influenced the design of the ELP has to do with diversity. Because individuals learn and demonstrate what they know in different ways, at different rates, and from the vantage point of their different experiences, teaching needs to utilize many approaches and provide a variety of assessment opportunities for students to demonstrate their knowledge and skills (Darling-Hammond, Ancess, & Falk, 1995; Falk, 1998a, 1998b; Falk, MacMurdy, & Darling-Hammond, 1995; Garcia & Pearson, 1994; Gardner, 1983; Kornhaber & Gardner, 1993; Price et al., 1993). The ELP is thus designed to collect multiple forms of evidence about a variety of forms of learning.

And finally, the ELP is designed to acknowledge the powerful role that interest and purpose play in motivating learning and in enabling students to show what they know (Arnold, 1995; Carini, 1986; Eisner, 1991; Perrone, 1991a). The ELP thus offers choice within its standardized format. For example, students participate in the selection of the text they read for their standardized reading interview; they write about topics of their own choosing when completing the standardized writing prompts; and they discuss issues of their own interest when being observed through the standardized format of the oral language assessment.

2. Literacy Learning and Effective Instruction

The ELP reflects a view of literacy informed by current reviews of literacy research (International Reading Association and the National Association for the Education of Young Children, 1998; National Research Council, 1998). It is based on the following assumptions:

Please do not cite, reproduce, or distribute without permission of authors

• Literacy involves four aspects of language use: reading, writing, listening and speaking. Each impacts on the other and influences the others' development.

• Literacy is an active process that involves obtaining meaning from and giving meaning to symbols - print. Literacy is about understanding the *world* as well as the *word*.

• Literacy learning is best conceptualized as a developmental continuum of progress rather than as an all-or-nothing phenomenon.

• Literacy learning is a multi-faceted process that requires experience and expertise with multiple factors. Effective teachers, therefore, utilize a mix of instructional ingredients crafted to suit the needs of each child.

• Early literacy learning is best supported by a balanced instructional approach that includes systematic guidance about the structure of language (alphabetic principle, phonemic awareness, phonics, and word recognition) as well as exposure to and immersion in rich literature and learning experiences.

• Children who experience difficulties in their literacy learning need the same rich literacy environment and mix of effective instructional ingredients as children who are progressing without difficulties. Struggling learners do not need different instruction than more able learners, they require *more intensive effective instruction* and *more intensive supports* to assist them.

• Accurate assessment of children's literacy knowledge, skills, strategies, and dispositions will help teachers better match instruction with how and what children are learning.

Guided by these principles, the ELP was designed to document students' abilities to:

Understand concepts about print: the overall structure of texts and conventions of the printed word (front/back of text, up/down and left/right directions of print, difference between individual letters and words)

Falk & Ort, 1999
Please do not cite, reproduce, or distribute without permission of authors

Use the three major cueing systems:
Graphophonic strategies - knowledge about written symbols of language - *phonemic awareness* (that speech is made up of different sounds), the *alphabetic principle* (that different sounds are represented by different symbols), and the ability to use these strategies for *word identification* (developing a substantial vocabulary that is recognized immediately and automatically)
Semantic strategies - context clues and prior knowledge/experience to recognize words and comprehend text
Syntactic cues - language structure and sentence grammar to recognize words and comprehend text

Comprehend: make sense out of print in order to summarize, sequence, analyze, interpret, predict, infer, and enjoy; monitor for understanding and to address misunderstandings.

Students' grasp of these essential literacy elements are evaluated in the Profile by examining the evidence of its tasks in relation to scales that describe stages of progress along a continuum of development. Profile design is intended to help teachers identify skills each student possesses and then place each student at a stage that best describes what the s/he knows and can do. Seeing the student in the context of the developmental continua is supposed to give teachers information that can be used to guide future instruction. It is also supposed to provide students and their families with a sense of where students are in the literacy process and what challenges they have to master in the future.

3. Principles For Reliable And Valid Assessment

To achieve the above aims, the ELP has incorporated into its design the following research-supported principles for reliable and valid assessment (Darling-Hammond & Falk, 1997a; Glaser & Silver, 1994; Linn et al., 1991; National Forum on Assessment, 1995; Valencia et al., 1994; Wiggins, 1993). The ELP aims to:

- Provide multiple forms of evidence about what students know, understand, and can do in many dimensions and kinds of learning

Please do not cite, reproduce, or distribute without permission of authors

Because learning is such a complex and variegated process, especially the process of literacy learning, relying on any one form of evidence to evaluate students' proficiencies and progress offers, at best, a limited view -- and sometimes even distorts the picture -- of what students actually know and can do. Multiple forms of evidence offer a more accurate picture of students' abilities (Price, Schwabacher, & Chittenden, 1993). Relying solely on one form of evidence for evaluating learning can be not only misleading but also harmful (Allington & McGill-Franzen, 1992; Darling-Hammond & Falk, 1997a; Falk, 1998a, 1998c; McGill-Franzen & Allington, 1993).

• Describe criteria for performance clearly and with detail
Assessments that are useful to learning provide accurate information about how students are progressing in relation to desired goals. Such assessments clearly and publicly articulate criteria for what students are expected to know and do in a particular discipline or area are so that they provide both teachers and students with a guide for learning (Darling-Hammond & Falk, 1997b; Falk & Ort, 1998; Herman, Aschbacher, & Winters, 1992; New York State Curriculum and Assessment Council, 1994; Resnick, 1994, 1995; Rothman, 1995, 1997).

• Measure the use of knowledge and skills embedded in meaningful, purposeful contexts and applications
Assessments that call on students to apply their knowledge in real-world situations and to demonstrate what they understand enable students with different learning styles and strengths to demonstrate their proficiencies in a variety of ways (Chittenden & Courtney, 1989; Darling-Hammond, Ancess, & Falk, 1995; Falk, 1998a; Falk & Larson, 1995; Falk, et al., 1995; McDonald, Smith, Turner, Finney, & Barton, 1993; Mitchell, 1992; Perrone, 1991b).

• Provide information that enhances teaching and supports learning
When assessments reveal the process as well as the product of learning, they help teachers shape their instruction in ways that are responsive to student needs. They encourage teachers to inquire and reflect -- about their students, about their discipline, and about their teaching strategies. In this way, they guide teachers, students, and their families to a better understanding of progress and growth. The assessment process thus becomes a learning experience for all members of the learning community (Darling-Hammond, Ancess, & Falk, 1995; Falk, 1994; Falk & Darling-Hammond, 1993; Falk & Ort, 1998; Shepard, 1995; Wolf, 1989; Wiggins, 1989; Wood & Einbender, 1995).

• Be accessible to students of diverse backgrounds
Assessment format and procedures need to be responsive to cultural, linguistic and regional differences. Flexibility in the response format

Please do not cite, reproduce, or distribute without permission of authors

allows students from diverse backgrounds and perspectives to demonstrate what they understand and what they can do.

• Reveal students' progress over time in relation to goals or standards for the discipline as well as in relation to reasonable expectations by age or developmental stage

Assessments that provide an indication of how students have progressed over time in relation to standards offer a clearer and more valid picture of achievement than those that focus only on outcomes without regard to students' starting points. Because students and groups of students may vary greatly in their levels of performance -- due to differences in family backgrounds and/or issues, physiological make-up, and/or language proficiencies -- assessment scores, to be most helpful, should indicate who started where and how far each has traveled in the journey toward proficiency. Measuring student progress in this way reveals and recognizes the value that teachers and schools have added to what students know and can do. This way of assessing promises to furnish a fairer picture of achievement than scores that simply provide information about how students compare to a national norm (Chittenden & Courtney, 1989; Falk, 1998b, 1998c; Falk & Darling-Hammond, 1993; Falk, MacMurdy & Darling-Hammond, 1995).

The Assessment Design Challenge: Serving Two Purposes

The ELP is designed to be a reliable and valid indicator of student progress that can be used for dual purposes: to inform teaching and support student learning as well as to report group performance trends. We have conducted studies of the Profile to ascertain how well the ELP meets these goals.

Much of the work in the field of assessment to date suggests that different types of assessments are designed to fulfill primarily different purposes -- some to furnish information that is useful for instruction, some to offer evidence of learning that is the result of a specific instructional experience, some to shed light on individual and group progress in order to address public questions about accountability. These different purposes require assessment forms that, in order to be technically sound, possess unique qualities and characteristics (Herman, Aschbacher, & Winters, 1992).

Please do not cite, reproduce, or distribute without permission of authors

Assessments that are to be used for reporting (or accountability) have to be standardized enough to mean the same thing in different places -- so that what is considered to be "accomplished" work in one locale represents the same level of accomplishment in another. Assessments used for accountability purposes also need to provide evidence that can translate into manageable and publicly-accessible information about the performance of students across locales and groups.

Assessments that are most useful for teaching however -- revealing what students know and can do, the strategies they use, their unique strengths, interests, and needs -- are, because of their very nature, difficult to standardize and translate into data that can reveal the performance trends of groups of students. (Student work samples, teacher observations, and/or projects that take place in the classroom are examples of these kinds of assessments.) Such assessments are difficult to standardize and translate into scores because they are highly sensitive to differences in classroom environments, local resources, and/or teachers' judgments. In addition, such assessments tap into complexities of subject matter and students' thinking that are difficult to measure and compare across groups. It is these very kinds of assessments, however, that reveal the richest picture of student knowledge and learning.

Herein lies a problem that is central to efforts to develop assessments that are useful for reporting and, at the same, are supportive of teaching and learning. Currently, to reliably look at student performance across groups, far too many assessments are standardized in a way that limits their abilities to provide instructionally useful information. The press for standardization often drives large-scale assessments to focus on what is easiest to measure rather than what is most important. As a result, these assessments generally measure lower-order skills in somewhat artificial formats. To make matters worse, the added pressure to do well on such tests, often drives instruction to directly mimic test content and format, sometimes creating conflict with worthy learning goals.

Falk & Ort, 1999
Please do not cite, reproduce, or distribute without permission of authors

This is the dilemma that the ELP has attempted to address: How to develop an accountability assessment that can reliably report on student progress while remaining valid and useful to teaching and learning. We have attempted to create an assessment that has the uniformity necessary to view progress across groups in trustworthy ways and yet also is sufficiently context-embedded and flexible so as to be responsive to individual differences, represent real-world performance, and capture students' genuine abilities and understandings (Linn, 1987; Moss, 1994; National Forum on Assessment, 1995). Our studies suggest that we have made significant progress toward meeting these goals. The data we present indicate that the ELP is a valid and reliable assessment instrument for monitoring and supporting individual progress as well as for reporting performance trends of large groups.

PART II: PROFILE STUDIES

To evaluate the ELP as a valid and reliable measure of literacy progress, a series of studies was conducted in 1997-1998 . Much of the data collected about the ELP was subjected to independent analysis by Katie Moirs of the Connecticut State Education Department to ensure technical accuracy and objectivity.

Pilot Sample

In 1997-98, 63 teachers representing 19 schools from 19 New York State school districts piloted the ELP with approximately 1215 students in grades 1-3. The sample was drawn to reflect the racial, socioeconomic, and regional diversity of the state, to represent the different types of locales in the state - small urban, large urban, suburban, and rural areas, and to include representation from the state's special needs and the linguistically diverse student populations. Although the sample was chosen from volunteers at the district level, the actual teachers who participated in the pilot were assigned.

Falk & Ort, 1999
Please do not cite, reproduce, or distribute without permission of authors

Description of the Studies

A brief summary of the studies conducted on the Profile follows:

1. **Construct Validity**: Does the ELP measure the intended trait(s) that are embodied in a definition of literacy? Does the ELP relate to other aspects of the domain of reading in the way that a theory of reading performance would predict?

To address issues of construct validity, the Profile was reviewed by a literacy assessment expert. A bias review was also conducted.

2. **Content Validity**: Is the ELP consistent with the curriculum it is a part of? Does it relate to the NYS standards?

To address issues of content validity, all teachers (n=63) who participated in the ELP pilot completed a survey consisting of a Likert scale and open-ended questions. Teachers were asked to evaluate the degree to which the ELP reflects the New York State English Language Arts Standards, matches teachers' curriculum, has positive effects on teachers' abilities to provide effective instruction relative to the standards, and correlates with teacher knowledge of students' literacy progress. Information such as race/ethnicity, gender, district type, years of teaching experience, class size, and educational background was also collected so that the degree to which these factors influenced teachers' responses could be assessed.

3. **Student Performance**: Is student performance on the ELP significantly differentiated by factors such as regional, racial, gender, socio-economic, linguistic diversity, or special education status?

To determine the degree to which student performance is differentiated by the above factors, the pilot sample was selected to represent the major geographic regions and types of locales in New York State - rural, suburban, small city, big city, and New York City. The total sample of students was also selected to include students enrolled in special education programs and those identified as Limited English Proficient (LEP), roughly in proportion to their representation in schools throughout the state. Demographic information such as race/ethnicity, gender, socio-economic status was collected for all students participating in the pilot so that scores could be analyzed based on these factors.

4. **Criterion Validity**: Does the ELP behave like other measures of this trait?

To address issues of criterion validity, all piloting third grade teachers (n=21) also administered several tasks to their students (n=363) from released items of a 4th grade NAEP reading and writing

Please do not cite, reproduce, or distribute without permission of authors

assessment. Tasks were scored by NCREST in consultation with associates from Educational Testing Service. Scores were correlated with ELP scale scores. The purpose of administering the NAEP assessment was not to compare the performance of individuals on the two exams but rather to compare trends of student performance.

We also collected student scores on the Degrees of Reading Power (DRP) test that, until 1999, was administered each spring to 3rd graders in New York State. These spring DRP scores were collected for 3rd graders participating in the pilot (n=289) and correlated with ELP scale scores.

5. **Reliability/Generalizability**: Can the ELP be scored reliably?

To address issues of reliability, we convened a summer scoring session with a universal sample (n=63) of the teachers involved in piloting the ELP. After this scoring session we convened a small group of expert scorers to blindly double score approximately 10% of randomly selected piloted ELP's. Using generalizability theory and the percent agreement method, interrater reliability for the ELP was estimated.

Out of the larger sample of selected papers, we randomly selected 20 completed ELP's to be scored by multiple "expert" scorers. Again, using generalizability theory, interrater reliability was estimated for these Profiles.

Findings

The pilot sample was analyzed for demographic and other variables. A presentation of the findings follows:

Teacher Demographics

Of the 63 teachers in the 1997-1998 pilot, 49% represented urban districts, 20% suburban, and 33% rural. 100% of the piloting teachers were women, 11% of whom were members of minority groups, and 89% of whom were white. We hypothesize that the strong representation of rural districts in the pilot sample impacted the distribution of white and minority teachers.

Teacher Experience

Most of the teachers in our sample had earned a Masters degree or higher: 76% had a Masters degree, 4% had PhD's or EdD's, and 17% had BA's. Teachers in our sample also tended to be quite experienced: 22% had between

Please do not cite, reproduce, or distribute without permission of authors

2-5 years experience, 33% 6-12 years, 22% had been teaching for 13-19 years, and 24% had more than 20 years experience. The average number of years of experience in our sample was 13, with a range of 2 to 32 years.

Class Size

The class size of the teachers who participated tended to be on the high end: 39% teach classes of 26-31+ students, 25% teach classes of 21-25, and only 26% have classes below 20 students. The average class size in our sample was 23, with a range of 15-31 students. Table I below summarizes the data about piloting teachers and districts:

TABLE I: PILOT TEACHER DATA

Teachers (n=46; Response rate =77%)	**Frequency**	**Percent**
Gender:		
Female	46	100%
Male	0	0
Ethnicity:		
Native American	1	2%
Latino/a	4	9%
White	41	89%
District Type:		
Urban	22	49%
Suburban	9	20%
Rural	15	33%
Highest Degree Earned:		
BA	8	17%
MA	35	76%
PhD/EdD	2	4%
Years Teaching: Range: 2-32 Mean: 13		
2-5	10	22%
6-12	15	33%
13-19	10	22%
20+	11	24%
Class Size: Range: 15-31 Mean: 23		
Below 20	12	26%
21-25	16	35%
26-31	18	39%

Falk & Ort, 1999
Please do not cite, reproduce, or distribute without permission of authors

BEST COPY AVAILABLE

Student Demographics

The 1215 students who participated in the 1997-1998 pilot of the ELP were fairly evenly distributed across the three piloting grade levels: 36% in grade 1, 33% in grade 2, 30% in grade 3.

Of the 1215 piloting students, 50% were boys and 50% were girls. Fifty-seven percent of students were white, 17% African American, 17% Latino/a, 5% Native American, and 4% Asian. Eight percent of piloting students received special education services, a number somewhat lower than overall statewide figures but representative of the numbers of students in special education in the early childhood grades. Eleven percent of students in the sample were identified as Limited English Proficient and 33% received compensatory or remediation services. The table below summarizes this data:

TABLE II: PILOT STUDENT DATA

	n	Frequency	Percent
Gender:	1215		
Female		611	50%
Male		604	50%
Ethnicity:	1146		
African American		195	17%
Latino/a		199	17%
Asian		44	4%
White		656	57%
Native American		52	5%
Special Education:	1215		
Yes		100	8%
No		1115	92%
LEP:	1215		
Yes		134	11%
No		1081	89%
Compensatory/ Remediation:			
Yes	1208	396	33%
No		812	67%
Free/ReducedLunch:	1212		
Yes		670	55%
No		542	45%
Grade Levels:	1215		
1		432	36%
2		400	33%
3		356	30%

BEST COPY AVAILABLE

Falk & Ort, 1999
Please do not cite, reproduce, or distribute without permission of authors

Summary of Data Related to Content Validity:

The construct review and teacher feedback were the main sources of data for assessing the ELP's content validity. The construct review evaluated the ELP to be true to the principles that guided its design. Teacher feedback from a survey administered to all piloting teachers constructed to ascertain their views about the overall relevance and effectiveness of the ELP, indicated that teachers overwhelmingly viewed the ELP positively. Very high percentages noted that the ELP provides valid and useful information about students' literacy progress; is useful to their teaching; allows a range of students to demonstrate what they know and can do; fits into the activities of their classrooms, and is useful for informing parents about their child's literacy progress.

Connections Between Curriculum and Assessment:

The ELP appears to be compatible with the teaching methods and strategies utilized in pilot teachers' classrooms. The vast majority of teachers (98%) reported that the ELP's ways of collecting evidence about literacy progress fit well with the teaching strategies and assessment methods that they currently provide for students. Many indicated that the Profile gave them a framework and systematic method for collecting the kind of student information that they need to effectively meet individual students' needs. One teacher commented:

> The activities in the ELP mirrored activities and assessment procedures already in place in my classroom. The Profile provided a more standardized method of reporting information.

Informing Parents about Students' Progress

At the same time that teachers felt the ELP fit well with their teaching strategies and assessment methods, 100% of teachers also indicated that the ELP has been useful for informing parents about their children's literacy progress. One teacher related her experience using the Profile with parents:

Falk & Ort, 1999
Please do not cite, reproduce, or distribute without permission of authors

> I was able to show a mother of a struggling child a clearer picture of his strengths and challenges and how he compares to required standards. As a result she is giving him more help at home. Parents need to know the guidelines and what level their children are at. This has given them a thorough picture.

Other teachers reported that the Profile has influenced changes in the ways their districts report student progress to parents:

> Pieces that I shared with parents were very well received. Our district report card committee is looking to the scales and summaries to change the language of our report card.

Manageability

Of the teachers surveyed, 80% reported that the ELP is "do-able" in a "reasonable" amount of time. As teachers' experience with the Profile increased, however, their perception of the "do-ableness" of the Profile also increased. Many teachers noted, in response to open ended survey questions, that as they became more familiar with the Profile, they were able to collect evidence more effectively and efficiently. They reported that the spring collection of evidence was significantly easier to do than the fall collection.

Usefulness to Teaching

Despite the 20% of teachers who expressed concern about time management in regard to administering the Profile, the majority of teachers surveyed reported that using the ELP was well worth the effort; 98% reported that the ELP provides information that is useful to their teaching. Typical responses to open-ended questions were:

> The ELP was useful to me in the sense that it helped me as a teacher to adjust my teaching techniques, to concentrate on some of the elements of literacy learning that I might have ignored, and to ask more penetrating questions to help the

Falk & Ort, 1999
Please do not cite, reproduce, or distribute without permission of authors

> students. The scales showed me specifics about where the students were as readers and writers.
>
> Observing the strategies a child uses and those which he doesn't helped me plan ways to help the child use those strategies he's not using now. I found myself planning with each student's needs in mind.
>
> Writing evidence clarified the conferences I had with each child. Reading responses gave me insight into the children's particular interests. It helped me guide them toward more challenging material.

Other survey responses indicate that teachers view the ELP as a valid measure of what constitutes literacy: 89% reported that the Profile adequately communicates students' literacy progress in relation to the NYS Standards for Learning in the English Language Arts; 87% reported that the ELP is a fair and accurate assessment that correlates with what they know about their students' learning. In response to open-ended questions, teachers made the following comments about the Profile's accuracy and validity:

> The scale scores correlated perfectly with what I know about my students' literacy progress. I found that the scores reflected what I had observed. The ELP gave me a more accurate and detailed description of where my students were at. I found that I had "proof" and evidence to support my observations.
>
> I was astonished at the accuracy of the rates of how highly the scores correlated with the students' abilities.
>
> The scores reflected what I already knew about the child and in a few cases helped me to realize that some students were weaker than I thought. Scale scores monitor growth and inform instruction.

Numerous teachers also noted how the Profile's information complements their other sources of information about student's literacy learning:

Falk & Ort, 1999
Please do not cite, reproduce, or distribute without permission of authors

> Scale scores closely correlate with students' progress judging from independent work, homework, and writer's notebooks.
>
> The results [Profile scores] validate and are validated by other reading/writing standards and assessments [we use with students].

Effectiveness as an Evaluation Tool

Survey responses provided insights to teachers' perceptions about the ELP's effectiveness as an assessment instrument in relation to other evaluation tools. 89% of teachers thought that the ELP more accurately and usefully measures literacy progress than traditional, norm referenced, multiple choice tests. 96% of teachers reported that they felt the ELP was effective in allowing a range of students to show what they know and can do in terms of literacy progress. 89% of teachers noted that they could confidently assign scale scores based on another teacher's collection of Profile evidence. And finally, 94% of piloting teachers indicated that the Profile scales are useful descriptions of the stages in the continuum of development in reading and writing.

Table III summarizes teacher responses on the survey:

Falk & Ort, 1999
Please do not cite, reproduce, or distribute without permission of authors

TABLE III: TEACHER EVALUATIONS

Survey Questions n = 46	Agree		Disagree		Don't know		No Response	
	n	%	n	%	n	%	n	%
Q1: The ways of collecting evidence found in the ELP resemble the kinds of activities that I provide for students in my classroom.	45	98%	0	0	0	0	1	2%
Q2: The ELP could be used to inform parents' understandings of their child's literacy progress.	46	100%	0	0	0	0	0	0
Q3: I found that collecting evidence for the ELP is do-able in a reasonable amount of time.	37	80%	8	17%	1	2%	0	0
Q4: The ELP provides information about my students' literacy progress that I can use in my teaching.	45	98%	0	0	0	0	1	2%
Q5: The ELP adequately communicates students' literacy progress vis a vis the NYS Standards for Learning in the English Language Arts.	41	89%	1	2%	4	9%	0	0
Q6: The ELP fairly and accurately assesses students' overall literacy progress.	40	87%	2	4%	4	9%	0	0
Q7: The ELP allows a range of students to show what they know and can do in terms of literacy progress.	44	96%	0	0	1	2%	1	2%
Q8: The ELP more accurately and usefully measures literacy progress than traditional, norm referenced, multiple choice tests.	41	89%	1	2%	4	9%	0	0
Q9: I could look at another teacher's ELP evidence and confidently assign his/her students a reading and writing scale score.	41	89%	2	4%	2	4%	1	2%
Q10: The ELP scales are useful descriptions of the stages in the continuum of development in reading and writing.	43	94%	2	4%	0	0	1	2%

BEST COPY AVAILABLE
Falk & Ort, 1999
Please do not cite, reproduce, or distribute without permission of authors

Teacher survey responses were also examined in relationship to a set of variables that were identified as possibly triggering differential response patterns. A one-way ANOVA was run using class size, school type (urban, rural, suburban), and years of teaching experience as independent variables. Aggregations of scaled survey responses were the dependent variable. No statistically significant differences based on these variables were found.

Summary of Student Performance Data

The ELP was administered to a sample of 1215 students in grades 1 - 3. Here we report overall mean scores, mean scores by grade levels, and frequencies of scores by grade levels. Overall mean scores for reading and writing were calculated based on an eight point scale, divided into four stages. Listening/Speaking scores were calculated based on a four stage scale.

Looking across all the students in the three grade levels involved in the pilot, the average reading score in the fall was 4.12 (beginning stage), the average writing score was 3.82 (beginning stage), and the average listening/speaking score was 2.49 (independent stage). By the spring, the average reading score increased to 4.69 (independent stage), the average writing score increased to 4.60 (independent stage), and the average listening/speaking score increased to 2.91 (independent stage). Table IV summarizes results pertaining to score means:

TABLE IV: OVERALL SCORE MEANS (based on 8 scale points)

	n	Min	Max	Mean	Std. Dev.
Fall Reading	1206	1	8	4.12	1.46
Fall Writing	1182	1	8	3.82	1.30
Fall Listening/Speaking	1169	1	4	2.49	.77
Spring Reading	1189	1	8	4.69	1.58
Spring Writing	1186	1	8	4.60	1.30
Spring Listening/Speaking	1190	1	4	2.91	.74

Please do not cite, reproduce, or distribute without permission of authors

Analysis of student performance on the ELP reveals that, for the most part, as students progress from grade to grade, on average their reading, writing, listening/speaking scores rise. In the *spring* of first grade, the average reading score was 3.91 (beginning stage), the average writing score was 3.73 (beginning stage), and the average listening/speaking score was 2.82 (independent stage). In the *spring* of second grade, the average reading score was 4.97 (independent stage), the average writing score was 4.81 (independent stage), and the average listening/speaking score was 2.95 (independent stage). In the *spring* of third grade, the average reading score was 5.41 (independent stage), the average writing score was 5.42 (independent stage), and the listening/speaking score was 3.00 (beginning stage). The Table V summarizes the data related to student scores by grade levels:

Falk & Ort, 1999
Please do not cite, reproduce, or distribute without permission of authors

TABLE V: STUDENT MEAN SCORES BY GRADE LEVELS
(based on 8 scale points)

	n	Min	Max	Mean	Std. Dev.
Grade 1					
Reading					
Fall	427	1	7	3.29	1.33
Spring	*421*	*1*	*7*	*3.91*	*1.49*
Writing					
Fall	405	1	6	2.90	1.08
Spring	*420*	*1*	*7*	*3.73*	*1.15*
Listening/Speaking					
Fall	411	1	4	2.33	.76
Spring	*421*	*1*	*4*	*2.82*	*.74*
Grade 2					
Reading					
Fall	399	1	8	4.35	1.43
Spring	*396*	*1*	*8*	*4.97*	*1.55*
Writing					
Fall	398	1	7	4.10	1.17
Spring	*394*	*2*	*8*	*4.81*	*1.07*
Listening/Speaking					
Fall	398	1	4	2.62	.84
Spring	*397*	*1*	*4*	*2.95*	*.79*
Grade 3					
Reading					
Fall	353	2	8	4.91	1.12
Spring	*348*	*2*	*8*	*5.41*	*1.26*
Writing					
Fall	352	2	8	5.60	1.06
Spring	*348*	*2*	*8*	*5.42*	*1.11*
Listening/Speaking					
Fall	333	1	4	2.56	.65
Spring	*348*	*1*	*4*	*3.00*	*.66*

Table VI below describes the score frequencies and percentages within each grade level for the four major stages of literacy progress identified by the ELP:

BEST COPY AVAILABLE

TABLE VI: SCORE FREQUENCIES BY GRADE LEVELS

Reading Scores (spring) n=1165	Grade 1		Grade 2		Grade 3	
	f	%	f	%	f	%
Emergent	79	18.8%	27	6.8%	3	.9%
Beginning	210	49.9%	115	29.0%	71	20.4%
Independent	111	26.4%	184	46.5%	213	61.2%
Experienced	21	5.0%	70	17.7%	61	17.5%

Writing Scores (spring) n=1162	Grade 1		Grade 2		Grade 3	
	f	%	f	%	f	%
Emergent	55	13.1%	4	1.0%	3	.9%
Beginning	271	64.5%	158	40.1%	56	16.1%
Independent	92	21.9%	209	53.0%	236	67.8%
Experienced	2	.5%	23	5.8%	53	15.2%

Listening/ Speaking Scores (spring) n=1166	Grade 1		Grade 2		Grade 3	
	f	%	f	%	f	%
Emergent	11	2.6%	12	3.0%	3	.9%
Beginning	129	30.6%	98	24.7%	65	18.7%
Independent	207	49.2%	184	46.3%	207	59.5%
Experienced	74	17.6%	103	25.9%	73	21.0%

These figures were used to determine reasonable expectations for stage/grade correlations:

Grade	Reasonable Expectations
First Grade	Beginning Stage (scale points 3 and 4)
Second Grade	Beginning/Independent Stages (scale points 4 and 5)
Third Grade	Independent Stage (scale points 5 and 6)

Student performance on the ELP was also examined in relationship to factors such as race/ethnicity, gender, limited English proficiency, socio-economic, compensatory/remedial education, and special education status. A

Please do not cite, reproduce, or distribute without permission of authors

one-way ANOVA was run using these factors as independent variables and student performance as the dependent variable. Statistically significant relationships were found for all independent variables.* Because differences in student performance based on these variables have been identified in the literature (Darling-Hammond, 1991, 1994; FairTest, 1999; Garcia & Pearson, 1994) and because teacher survey responses of the Profile as well as the expert bias review found it to be sensitive to issues of diversity, we hypothesize that the differential student performance may be due to inequities in resources and opportunities to learn for students from different racial/ethnic, socio-economic, and linguistic backgrounds. Lower student performance from poor and minority communities may also reflect the fact that these communities have less access to qualified teachers and quality materials (National Commission on Teaching and America's Future, 1997).

Summary of Correlational Data

As part of the studies conducted, we correlated the ELP scores of third grade pilot students with their scores on other measures of literacy progress. Table VII presents correlations of fall and spring reading and writing Profile scores with student scores on released items from a 4th grade NAEP reading and writing assessment and student scores on the Degrees of Reading Power (DRP) test. The NAEP tasks generated separate scores for reading, writing, and multiple choice items. Profile scores generated by one scorer, the classroom teacher, were used in this analysis.

* Authors will provide more detailed information upon request.

Please do not cite, reproduce, or distribute without permission of authors

TABLE VII: ELP SCORES CORRELATED WITH NAEP AND DRP SCORES

	Reading Fall, 1997	Writing Fall, 1997	Reading Spring, 1998	Writing Spring, 1998
NAEP Writing Scores	.443**	.336**	.489**	.453**
NAEP Reading Scores	.320**	.176**	.353**	.313**
NAEP M/C Scores	.207**	.145*	.265**	.299**
DRP Scores	.577**	.375**	.605**	.536**

* $p \leq .05$;
** $p \leq .01$

As shown in Table VII, DRP and NAEP scores have statistically significant, but relatively moderate correlations with ELP scores. Of all the correlations, the highest is between spring reading Profile scores and DRP scores ($r = .605$). From a construct validity perspective, these correlational findings make sense. Both the DRP and the NAEP tests should be measuring in part what the ELP is measuring, so both should be somewhat correlated with Profile scores. However, in this case only moderate correlations are desired. A performance assessment should be capturing a complexity that paper-and-pencil instruments are not capable of measuring. The ELP assessment should be measuring something unique: achievement not demonstrated through performance on either the NAEP or DRP tests due to the limitations of these tests as measures of certain types of achievement. Correlation study findings support this theory, suggesting that while the ELP is measuring achievement common to that measured by both the NAEP and DRP tests, it is also measuring achievement that can be captured only in the performance opportunities that are unique to the ELP.

Summary of Interrater Reliability Data

Our analyses indicate that the ELP can be scored reliably. When the Profile is scored by two scorers, a interrater reliability exceeding .8 is achieved.

Please do not cite, reproduce, or distribute without permission of authors

To determine reliability coefficients, approximately 10% of the ELP's were blindly double-scored by teachers involved in piloting the instrument. At scoring sessions in the fall and the spring, piloting teachers were asked to count off every eighth completed and scored Profile from amongst their classroom sets. These Profiles were subsequently collected, organized, coded, and assigned into piles for rescoring. In this manner, we were able to ensure that each set of Profiles to be rescored represented all of the teachers and students in the pilot. We were also able to ensure pilot teacher and student anonymity. The Profiles were then rescored by a specially trained group of teachers who were selected based on their experience using the ELP and their expertise in literacy assessment. In the same manner that all of the piloting teachers were trained, a protocol was conducted for the expert scorers that instructed them to 1) examine the evidence, 2) choose one of the four major stages of the ELP scales that best described the evidence, and then 3) assign a scale point score that more finely described the qualities evident in the student work.

Using scores generated by this data collection procedure, interrater reliability for the ELP was estimated using two methods: generalizability theory and percent agreement.

Generalizability Theory

Generalizability theory provides the most flexible and useful approach for estimating interrater reliability for performance assessments such as the ELP. While classical test theory postulates that an observed score can be decomposed into a true score and an error term, generalizability theory uses analysis of variance techniques to disentangle the error term into multiple sources. G-study components yield estimates for all sources of variance included in a particular design (e.g., rater-by-task), which are then used in a D-study for estimating variance components over an increasing number of raters. These D-study variance components are used to estimate variances

Please do not cite, reproduce, or distribute without permission of authors

and reliability-like coefficients, referred to as generalizability coefficients, which represent ratios of universe to observed score variance.

Table VIII also shows the G-study variance components and the D-study generalizability coefficients generated by generalizability analyses using both sets of ELP reading and writing scores (fall and spring) on the 4 stage scale. Analyses were based on a Profile (person)-by-scorer (rater) design. In this design, the Profile score is the independent variable, and G-study variance component estimates include: Profile, scorer, and the interaction component of Profile-by-scorer. The Profile component is an estimate of the variance across Profiles of Profile level mean scores where the mean is taken across all scorers in the universe. This variance component should be greater than zero, indicating that scorers were able to differentiate between different levels of student performance. The scorer component is an estimate of the variance of scorer mean scores, where each mean is taken across Profiles. A scorer variance estimate close to 0, or relatively small, indicates that the scorer facet does not contribute, or contributes very little, to score variability. Likewise, a close to zero or relatively small Profile-by-scorer variance component suggests that the various scorers were able to rank order evidence sets similarly.

Please do not cite, reproduce, or distribute without permission of authors

TABLE VIII: INTERRATER RELIABILITY ANALYSES SUMMARY TWO SCORERS (4 stage scale)

Generalizability Analyses

	Reading Scores Fall (n=149)	Writing Scores Fall (n=149)	Reading Scores Spring (n=157)	Writing Scores Spring (n=156)
Variance Components (G-study):				
Profile (p)	.378	.33	.45	.34
Scorer (s)	-.00	.01	-0.00	.00
Profile X Scorer (ps)	.14	.16	.21	.12
Generalizability Coefficients (D-study):				
1 scorer	.74	.68	.68	.73
2 scorers	.85	.81	.81	.85
3 scorers	.89	.86	.86	.89
4 scorers	.92	.89	.89	.92
5 scorers	.93	.91	.91	.93
6 scorers	.94	.93	.93	.94
7 scorers	.95	.94	.94	.95
8 scorers	.96	.94	.94	.96
9 scorers	.96	.95	.95	.96

Perfect and Adjacent Agreement Statistics

	Reading Scores Fall (n=149)	Writing Scores Fall (n=149)	Reading Scores Spring (n=157)	Writing Scores Spring (n=156)
Perfect Agreement	109 (73%)	100 (67%)	103 (66%)	117 (75%)
Adjacent Agreement	40 (27%)	49 (33%)	50 (32%)	39 (25%)
2 points off			4 (2%)	
3 points off				

As shown by Table VIII, variance estimates for both sets of reading and writing scores (fall and spring), for both scales, are relatively small, with the largest variance component for each analysis being the main effect for the Profile. Scorer and scorer-by-Profile variance estimates all close to 0, or relatively small, indicate that, in this study, the scorer facet contributed little to Profile score variability, while the relatively large Profile variance component indicates that the scorer pairs were able to differentiate between different levels of student performance on the ELP. In other words,

Please do not cite, reproduce, or distribute without permission of authors

BEST COPY AVAILABLE

differences in Profile scores were found to be related to the different performance levels reflected in Profiles and were not attributable to differences among the scorers.

The accompanying D-study identified the number of scorers per Profile that would be required to obtain acceptably small error variances or acceptably large generalizability coefficients. Table VIII also shows reliability estimates based on 1, 2, 3, 4, 5, 6, 7, 8, and 9 scorers for both sets of reading and writing scores (fall and spring). These estimates suggest that, based on the performance of scorers used in this study, using one scorer to score an ELP would result in generalizability coefficients falling close to the .7 range, while the use of two scorers to score the same ELP would result in generalizability coefficients falling within the .80 range or higher. Although the generalizability coefficients become larger as the number of scorers increases (e.g., 4 scorers would yield reliability estimates of .89 or higher), the number of scorers necessary to achieve acceptable levels of reliability must be considered in terms of feasibility.

Interrater Reliability Estimates Based on Five Scorers

To provide more information about interrater reliability, another type of generalizability study was conducted on a subset of ELP's. A small subset of the scored ELP's was randomly selected from amongst the randomly selected 10% of Profiles that were double-scored in the Spring. These were rescored by additional scorers. In the original design, five different scorers were to score an additional 20 Profiles in the spring. Final data collection looked like this:

Reading, Spring 1998	5 scorers, 18 Profiles scored
Writing, Spring 1998	5 scorers, 17 Profiles scored

Table IX shows results for the generalizability analyses using scores generated by the above scoring scheme. Results are shown for spring reading and writing scores for the 4-stage scales. These statistics are equally as promising as those obtained from analyses using Profile scores generated by

Please do not cite, reproduce, or distribute without permission of authors

two scorers. The five spring scorers were either in perfect or adjacent agreement for the writing scores (100% perfect or adjacent agreement, 17 Profiles scored), with 95% of the reading scores being either in perfect agreement or within one point. Further, generalizability analyses revealed a pattern of G-study variance components similar to that for the two-scorer sample discussed above, with solid generalizability coefficients for both sets of reading and writing scores.

TABLE IX: INTERRATER RELIABILITY ANALYSES SUMMARY
FIVE SCORERS

Generalizability Analyses

	Reading Scores - Spring (raters 1, 2, 3, 4 and 5) (n =18)	Writing Scores - Spring (raters 1, 2, 3, 4 and 5) (n = 17)
Variance Components (G-study):		
Profile (p)	.40	.20
Scorer (s)	-0.01	-0.00
Profile X Scorer (ps)	.18	.11
Generalizability Coefficients (D-study):		
1 scorer	.69	.65
2 scorers	.82	.79
3 scorers	.87	.85
4 scorers	.90	.88
5 scorers	.92	.90
6 scorers	.93	.92
7 scorers	.94	.93
8 scorers	.95	.94
9 scorers	.95	.94

Perfect and Adjacent Agreement Statistics (4 stage scale)

	Reading Scores - Spring (raters 1, 2, 3, 4 and 5) (n=18)	Writing Scores - Spring (raters 1, 2, 3, 4 and 5) (n=17)
Perfect Agreement	7 (40%)	11 (65%)
Adjacent Agreement	10 (55%)	6 (35%)
2 points off	1 (5%)	

Percent Agreement Method

The percent agreement method is simply an estimate of the degree to which Profiles that are independently scored by two or more different scorers

Falk & Ort, 1999
Please do not cite, reproduce, or distribute without permission of authors

BEST COPY AVAILABLE

agree across Profile scores. Perfect agreement is the percent of Profiles that are scored exactly the same by two or more scorers; adjacent agreement the percent scored plus or minus one point apart by two or more scorers.

Table VIII (p. 30) shows the perfect agreement and adjacent agreement statistics for both sets of reading and writing scores (fall and spring) based on independent scorings of the same Profile by two scorers. These statistics suggest that the scorer pairs scored the same Profile similarly. For both reading and writing scores (fall and spring), 100% of the pairs of Profile scores were either in perfect agreement or off by one point only.

Summary of Reliability Data

In summary, as indicated by two methods of estimating interrater reliability, generalizability theory and percent agreement method, scorers made judgments across ELP's with a high degree of reliability when rigorously trained according to a thorough and well-crafted scoring system. Further, trained scorers were able to differentiate among levels of student performance using the scoring system that has been developed for the ELP.

Discussion

The questions this study aimed to address are: Can a classroom-based, context-embedded assessment reliably and validly serve two purposes? Can it provide information about individual student progress as well as information about group performance trends that is useful to teaching and learning and can be used for reporting purposes? The findings presented herein suggest that this is indeed possible. We found the ELP to be a valid assessment of literacy progress that is **technically strong** - it effectively differentiates levels of student performance and can be scored reliably - and that is **instructionally useful**.

Instructional Usefulness

Perhaps the most powerful finding of our studies is the degree to which teachers reported the ELP to be supportive of their teaching and their

Please do not cite, reproduce, or distribute without permission of authors

students' learning. High percentages of the teachers reported that the Profile provides fair and accurate descriptions of children's literacy progress, yields information that is useful to instruction, connects to the New York State Standards for the English Language Arts, is reflective of the kinds of activities they provide in their classrooms, and enhances parents' knowledge of their children's progress. This affirmation by the teachers is especially significant in light of the amount of work that the Profile requires.

Both teacher survey responses and the construct review suggest that the ELP is instructionally useful. The data point to several aspects of the Profile that support student and teacher learning. Student learning is supported by Profile use because it is made up of tasks that embody the important learning goals expressed in the New York State Standards and because these tasks call on students to apply their skills and understandings in ways that are much like real-world performance. Teacher learning is also supported by the Profile in several ways. By virtue of what the Profile asks teachers to observe and record - critical literacy skills and behaviors - teachers are provided with a guide to essential aspects of the literacy learning process. In addition, Profile use leads teachers to collect and rely on authentic student work as evidence on which they can base instructional decisions. The Profile also gives teachers immediate feedback that can be instructionally helpful. Instead of having to wait for months after test administration to receive students' scores from the test publisher, district, or state (as is the case with many tests currently used), the ELP is designed for teachers to be the primary assessors and to have on-site, immediate access to information about their students' performance and progress.

The findings of our studies suggest that by asking teachers to look at evidence of student learning (as it is manifested in student work) in relation to standards (as described in the Profile scales), teachers perceive themselves to have increased their knowledge of individual students, to have become better informed about the capacities of their students in relation to literacy

Falk & Ort, 1999
Please do not cite, reproduce, or distribute without permission of authors

progress, and to have received guidance about what they need to do next to support the forward development of their students.

Based on these findings, we predict that Profile use over time will help teachers become even better informed about literacy. Not only will this enhance their overall pedagogical capacity, it will help to bring control of assessment back into their hands, away from the "outside experts" and commercial testing companies that presently dominate the assessment process. National studies and reports have documented the strong relationship between teacher quality and student performance. As a profession we now have data to demonstrate that increased professional knowledge on the part of teachers yields higher levels of student performance (National Commission on Teaching and America's Future, 1998). This study of the ELP leads us to predict that as teachers become more expert about literacy instruction due, in part, to what they learn from using assessments such as the ELP, we can expect to witness improved progress and performance on the part of their students.

Technical Strength

The technical merit of the ELP is also demonstrated by study results. Our findings indicate that the Profile is able to accurately describe literacy progress and differentiate student performance at different stages of development. In addition, the evidence suggests that the stages of development defined by the Profile are broadly related to different grade levels and that teachers' decisions about performance translate to scores that are consistent (reliable) across different scorers.

The construct review and teacher survey responses suggest that the Profile describes and assesses important dimensions of literacy. The correlations between student Profile scores and student scores on other measures of literacy (DRP and/or NAEP) provide further support. However, these correlations with other literacy measures, while affirming the ELP's construct validity, lead to other questions. If correlations are substantial, why

Please do not cite, reproduce, or distribute without permission of authors

should we consider using the more labor intensive, complex Profile instead of less time-consuming and easier to score existing tests? Our answer to this question is that while the correlational data indicate that the ELP reflects some of the same aspects of literacy revealed by the DRP and NAEP tests, the ELP also allows students to demonstrate some aspects of literacy that are difficult to capture in timed, predominantly multiple-choice tests. More importantly, the ELP is a preferred format, according to surveyed teachers, because they perceive it to be more instructionally useful.

Another indicator of the ELP's technical strength revealed by these studies is that score distributions among different grade levels reveal patterns of progress toward higher scale scores as students advance in the grades. This finding suggests that the Profile has the capacity to differentiate performance as might be expected for children of increasing ages. Score overlap at different grade levels (see Table VI) leads us to postulate some "Reasonable Expectations" for Profile stage acquisition in relation to grade level (see page 25). We advise however that the "Reasonable Expectation" framework be used only as a general guide rather than as a strict requirement that must be met by grade completion. Consistent with theories of human development, which postulate that children progress unevenly in their learning - in different ways and at different paces - we caution users of the ELP not only to expect and support individual variation in student performance, but to inform decisions about students' instruction with the broadest possible range of evidence.

Findings based on analyses of the ELP's interrater reliability indicate that it can be scored with consistency across raters and that rater judgments are reliable. The reliability rate exceeding .8 between two scorers resulting from generalizability theory analyses suggests that the Profile can be operationalized for reporting purposes. This finding indicates, however, that reliability will be strongest if two scorers examine and score each Profile. Our experience working with Profile pilots in districts throughout New York State leads us to suggest that double scoring is not only feasible but has professional

Falk & Ort, 1999
Please do not cite, reproduce, or distribute without permission of authors

development benefits. Double scoring can be performed in professional development half-days, after-school sessions, or summer institutes. Although providing for such sessions presents school and district administrators with a challenge to their time and fiscal resources, there are benefits to bringing teachers together to assess student work in relation to standards. These benefits have been documented in a several studies (Allen, 1998; Falk & Ort, 1998) which suggest that looking at student work in relation to standards strengthens teachers' understandings of their discipline, deepens their knowledge of their students, provides insights to teaching strategies, and enhances their sense of professionalism.

Conclusion

Overall, the findings from the studies conducted on the ELP are promising. They indicate that the ELP provides valid and useful information about student progress that, under the appropriate scoring conditions, can be used for reporting purposes. There are questions, however, that remain unanswered and that warrant further inquiry. In particular, future studies of the ELP might include the following: Does the ELP accurately predict student performance on subsequent measures of theoretically related traits? What are the long-term consequences of Profile use on teacher practice? Does it help teachers to teach better? Does it improve student learning? What issues emerge related to wide-scale implementation?

Educators and assessment experts continue to debate whether it is possible for large scale assessment to serve reporting purposes as well as to provide instructionally useful information to further student learning. The Early Literacy Profile was designed to contribute to the conversation about how to meet both of these very important functions. It is our hope that findings from this study demonstrate that one possible way of meeting these needs is to embed assessment into classroom life and involve teachers in scoring processes. Because, in the final analysis:

Please do not cite, reproduce, or distribute without permission of authors

Teachers, not assessments, must be the cornerstone of any systemic reform directed at improving our schools..."The teacher is a mediator between the knower and the known, between the learner and the subject to be learned. A teacher, not some {test}, is the living link in the epistemological chain" (George Madaus, quoting Parker Palmer, A National Testing System, 1992, p. 5).

Falk & Ort, 1999
Please do not cite, reproduce, or distribute without permission of authors

REFERENCES

Allington, R. L. & McGill-Franzen, A. (1992). Unintended effects of educational reform in New York. *Educational Policy, 6,* 397-414.

Arnold, K. (1995). *Lives of Promise: What becomes of high school valedictorians.* San Francisco: Jossey-Bass.

Barrs, M., Ellis, S., Hester, H., & Thomas, A. (1988). *The Primary Language Record.* London: ILEA/Centre for Language in Primary Education.

Bruner, J. (1960). *The process of education.* Cambridge, MA: Harvard University Press.

Carini, P. (1986). Building from children's strengths. *Journal of Education, 168,* 13-24.

Chittenden, E. & Courtney, R. (1989). Assessment of young children's reading: Documentation as an alternative to testing. In D. S. Strickland & L. M. Morrow (Eds.), *Emerging literacy: Young children learn to read and write* (pp. 107-120). Newark, DE: International Reading Association.

Darling-Hammond, L. (1991). The implications of testing policy for educational quality and equality. *Phi Delta Kappan, 73* , 220-225.

Darling-Hammond, L. (1994). Performance-based assessment and educational equity. *Harvard Educational Review, 54* , 5-30.

Darling-Hammond, L. (1997). *The right to learn: A blueprint for school reform.* New York: Jossey-Bass.

Darling-Hammond, L., Ancess, J., & Falk, B. (1995). *Authentic assessment in action.* New York: Teachers College Press.

Darling-Hammond, L. & Falk, B. (1997a). Policy for authentic assessment. In A. Lin Goodwin (Ed.). *Assessment for equity and inclusion.* London: Routledge.

Darling-Hammond, L. & Falk, B. (1997b). *Using standards and assessments to support student learning. Phi Delta Kappan , 79* , 190-199.

Education Department of Western Australia. (1994). *First Steps.* Melbourne, Australia: Longman.

Eisner, E. W. (1991). What really counts in schools. *Educational Leadership, 48,* 10-17.

Please do not cite, reproduce, or distribute without permission of authors

FairTest. (1999). California to reduce SAT emphasis. *Examiner, 13*(1), p. 1 & 4. Cambridge, Mass.: Author.

Falk, B. (1994). *The Bronx New School: Weaving assessment into the fabric of teaching and learning.* New York: National Center for Restructuring Education, Schools, and Teaching (NCREST).

Falk, B. (1996). Teaching the way children learn. In M. McLaughlin and I. Oberman (Eds.). *Teacher learning.* New York: Teacher College Press.

Falk, B. (1998a). Testing the way children learn: Principles for valid literacy assessments, *Language Arts, 76*(1), 57-66.

Falk, B. (1998b). Looking at students and their work: Supporting diverse learners with the *Primary Language Record*." In D. Allen (ed.), *Assessing Student Learning: From Grading to Understanding.* New York: Teachers College Press.

Falk, B. (1998c). Using direct evidence to assess student progress: How *The Primary Language Record* supports teaching and learning." In T. Salinger (ed.), *Assessing Reading: Theory and Practice.* London: Routledge.

Falk, B. & Darling-Hammond, L. (1993). *The Primary Language Record at P.S. 261: How assessment transforms teaching and learning.* New York: National Center for Restructuring Education, Schools, and Teaching (NCREST).

Falk, B. & Larson, J. (1995). *An invitation to invention: Top-down support for bottom-up reform of assessment in New York State.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Falk, B., MacMurdy, S., & Darling-Hammond, L. (1995). *Taking a different look: How the Primary Language Record supports teaching for diverse learners.* New York: National Center for Restructuring Education, Schools, and Teaching (NCREST).

Falk, B. & Ort, S. (September, 1998). Sitting down to score: Teacher learning through assessment. *Phi Delta Kappan.*

Fosnot, C. T. (1989). *Enquiring teachers, enquiring learners.* New York: Teachers College Press.

Please do not cite, reproduce, or distribute without permission of authors

Garcia, G.E. & Pearson, P. D. (1994). Assessment and diversity. In L. Darling-Hammond (Ed.), *Review of Research in Education, 20* (pp. 337-391). Washington, DC: American Educational Research Association.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences.* New York: Basic Books.

Glaser, R. & Silver, E. (1994). Assessment, testing and instruction: Retrospect and prospect. In L. Darling-Hammond (Ed.), *Review of Research in Education, 20* (pp. 393-419). Washington, DC: American Educational Research Association.

Griffen, P., Smith, P., & Burrill, L. (1995). *The American ELP scale: A framework for authentic assessment.* Portsmouth, NH: Heinemann.

Herman, J., Aschbacher, P., & Winters, L. (1992). *A practical guide to alternative assessment.* Alexandria, VA: ASCD.

International Reading Association and the National Association for the Education of Young Children. (1998). *Learning to Read and Write: Developmentally Apropriate Practices for Young Children.* Washington, D.C.: Authors.

Kornhaber, M. & Gardner, H. (1993). *Varieties of excellence: Identifying and assessing children's talents.* New York: National Center for Restructuring Education, Schools, and Teaching (NCREST).

Linn, R.L. (1987). Accountability: The comparison of educational systems and the quality of test results. *Educational Policy, 1,* 181-198.

Linn, R.L., Baker, E.L., & Dunbar, S.B. (1991). Complex, performance-based assessments: Expectations and validation criteria. *Educational Researcher, 20,* 15-21.

McDonald, J., Smith, S., Turner, D., Finney, M., & Barton, E. (1993). *Graduation by exhibition.* Alexandra, VA: Association for Supervision and Curriculum Development.

Madaus, G.F. (1992). *A national testing system: Manna from above: A historical/technical perspective.* Chestnut Hill, MA: Boston College, Center for Study of Testing, Evaluation, and Educational Policy.

Mitchell, R. (1992). *Testing for learning: How new approaches to evaluation can improve American schools.* New York: The Free Press.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher 23,* 5-12.

Please do not cite, reproduce, or distribute without permission of authors

National Association for the Education of Young Children. (January, 1988). NAEYC position statement on developmentally appropriate practice in the primary grades, serving 5 through 8 year olds. *Young Children*, 64-84.

National Commission on Teaching and America's Future. (1997). *Doing what matter most: Investing in quality teaching*. New York: Author.

National Forum on Assessment. (1995). *Principles and indicators for student assessment systems*. Cambridge, MA.: Fairtest.

National Research Council. (1998). *Preventing Reading Difficulties in Young Children*. Washington D.C.: National Academy Press.

New York State Curriculum and Assessment Council. (1994). *Learning-centered curriculum and assessment for New York State*. Albany, NY: New York State Education Department.

Perrone, V. (1991a). *A letter to teachers: Reflections on schooling and the art of teaching*. San Francisco: Jossey-Bass.

Perrone, V. (Ed.). (1991b). *Expanding student assessment*. Alexandria, VA: ASCD.

Piaget, J., & Inhelder, B. (1970). *The science of education and the psychology of the child*. New York: Penguin Books.

Price, J., Schwabacher, S. & Chittenden, E. (1993). *The multiple forms of evidence study*. New York: The National Center for Restructuring Education, Schools, and Teaching (NCREST).

Resnick, L. B. (1987). *Education and learning to think*. Washington, DC: National Academy Press.

Resnick, L.B. (1994). Performance puzzles. *American Journal of Education 102*, 511-526.

Resnick, L. (1995). Standards for education. In D. Ravitch (Ed.), *Debating the future of American standards*. Washington, DC: The Brookings Institution.

Rochester, New York Public Schools. (199?). *The student outcomes and developmental stages of the Rochester, New York Public Schools*. Rochester, NY: Author.

Falk & Ort, 1999
Please do not cite, reproduce, or distribute without permission of authors

Rothman, R. (1997). *Organizing schools so that all children can learn.* New York: National Center on Education and the Economy.

Rothman, R. (1995). *Measuring up: Standards, assessment, and school reform.* San Francisco, CA: Jossey-Bass.

South Brunswick, New Jersey Public Schools. (1992). *The K-2 reading/writing scale.* South Brunswick, NJ: Author.

Sternberg, R.J. (1985). *Beyond IQ.* New York: Cambridge University Press.

Valencia, S., Hiebert, E., & Afflerbach, P. (1994). *Authentic reading assessment: Practices and possibilities.* Newark, DE: International Reading Association.

Vygotsky, L. S. (1978). *Mind in society.* Cambridge, MA: Harvard University Press.

Wiggins, G. (1993). *Assessing student performance: Exploring the purpose and limits of testing.* San Francisco: Jossey-Bass.

Falk & Ort, 1999
Please do not cite, reproduce, or distribute without permission of authors